Research Paper | Computer Science | E-ISSN No : 2454-9916 | Volume : 6 | Issue : 5 | May 2020

VARIOUS SPAMS AND CLASSIFICATION ALGORITHMS FOR DETECTION OF SPAM EMAIL THROUGH COMPARING J48, SVM AND NAIVE BAYES CLASSIFIERS USING WEKA TOOL

Vutharkar Nagaveni [1] | Dr. Vimal Pandya [2]

[1] Computer Science, Rai University, Saroda, Gujarat, India.

[2] Director, Computer Science, Navgujarat College of Computer Applications, Ahmedabad, Gujarat, India.

ABSTRACT

Electronic Mail (E-mail) is playing most important and significant role taken in the world of information communication for users. Nowadays, Email is most common and effective mode of communication technology for communicate and sharing the information to both end users. The rapid increase of email users there will be increase of volume of spam emails too from the past few decades. Emails are categorized in to ham and spam emails. This paper illustrates on different existing email spam filter system regarding Machine Learning Technique (MLT) such as Naive Bayes, SVM, J48 and present the classifications, evaluation and comparison on different email spam filtering algorithms using WEKA Software and performs various parameters like finding Accuracy, Recall, Precision, Measures and False Position Rate etc. The final output result should be '1' if it is finally spam present, otherwise, it should be '0' for non-spam. In this analysis the Final out presents that J48 classifier is best and efficient algorithm for spam or not spam emails among other algorithms.

KEY WORDS: WEKA, Support vector Machine, Email, J48, Naive Bayes, Spam.

1. INTRODUCTION:

Emails are categories into two types Spam emails and Ham emails. Spam emails are the junk mails received from illegitimate users that contain advertisement, malicious code, virus or those who gain personal profit from the end user. Spam can be transmits from any source like Web, Text messages, Fax etc., depends on the mode of transmission, spam can be categorised into various categories like email spam, web spam, text spam, social networking spam etc. Now a days, spam contain viruses like Trojan horses and other harmful software that cause to failures the computer systems, networks, bandwidth and storage space and slows down email servers etc.

Now a days, spam mails needs anti-spam filters that are reliable, accurate, and effective for classification spam emails. There are several text mining, Data mining and Machine-Learning Techniques (MLT) available to classify spam mail such as Naïve Bayes, and Support Vector Machines, J48 Classifier, Bayes net classification and etc.

I used Machine learning algorithms for classification of objects in different classes to provide efficient classifying emails as spam or harm. In this research I, used three main machine learning algorithms namely, Naive Bayes classifier, Support Vector Classifier and J48 Classifier for classification of spam emails.

WEKA tool kit is a free and open-source software that compiles data-mining algorithms in machine-learning applications. WEKA also perform tasks like pre-processing, statistical processing and visualization of data etc., (www.cs.-waikato.ac.nz/ml/weka) and algorithms such as Naive Bayes classifier, Support Vector Classifier and J48 Classifier are applied to classify spam mail detection . The descriptions of these algorithms and comparison of their performance using the WEKA environment provides the final report analysis.

I have given a short review detail description of the three classification algorithms and present experimental details followed by results and discussion is provided. The Final conclusions followed by avenues for future work is also availed.

2. RELATED WORK:

In this paper author describe spam and Email spam with behaviour of characters of spam mail and various machine learning and non-machine learning algorithms. Various techniques to detect spam emails has discussed by the author in this survey. [1]

In this paper author describe as assessment of case base reasoning approach for long text message to short text message. In this evaluation it determines appropriate feature types and feature representation of short text messages uses Naïve Bayes classifier and support vector machine algorithms. [7]

In this study author explained various spam and its types with focus on Image Spam detection causing problems till now with all the solutions that have been developed by various venders and users causes. It still poses a great threat and still penetrate to the user's e-mail. [4]

In another research author discussed tremendous growing problem of phishing e-mail, also known as spam including spear phishing or spam borne malware, has demanded a need for reliable intelligent anti-spam e-mail filters. This survey work presented and discussed the types and implication of spam emails on modern society and commerce. This survey paper focused on literature survey of Artificial Intelligence (AI) and Machine Learning (ML) methods for intelligent spam email detection, which helps in developing appropriate counter measures. [3]

In this research work modified J48 classifier has been used to increase the accuracy rate of the data mining procedure with new approach for efficiently predicting the diabetes from medical records of patient the data mining tool WEKA has been used as an API of MATLAB for generating the modified J-48 classifiers. Experimental results shown a significant improvement over the existing J-48 algorithm. It is proved that the proposed algorithm can achieve accuracy up to 99.87 %. [12]

3. DESCRIPTION OF SPAM AND TYPES OF SPAM:

3.1. Spam:

Spam is commonly defined as unsolicited bulk email messages received without one's permissions, and the goal of spam detection is to distinguish between spam and legitimate email messages. Most of the spam contain viruses, Trojan horses and other harmful software that may cause to failures computer systems, networks, and bandwidth and storage space to slow down email servers.

Spammers collects email IDs from various sources such as chats, websites, newsgroups , malware and address details of users, which are easily available from other spammers for low price and bulk of messages are sent to recipients where, the volumes of which create enormous productivity losses to IT firms and huge serious security threats that carriers classified information. Hence, the classification of emails is prime importance to handle spam emails.

3.2 Email Spam:

Internet is important and essential part of human life. Email is the simplest and fastest mode of communication over the internet that is used both personally and professionally. Due to the increase in the number of account holders and increase in the rate of transmission of emails as a serious issue of spam emails around the world. An survey it was analysed that over 294 billion emails are sent and received every day and found that Over 90% emails are reported to be spam emails .

The rate at which email spamming is spreading is increasing tremendously because of fast and immodest way of sharing information. It was reported that user receives more spam mails than ham mails. Spam filtration is important because spam waste time, energy, bandwidth, storage and consume other resources.

Email can be categorised as a spam email by the following characteristics:

1. ***Unsolicited Email:*** Email that was received by unknown contact or illegitimate contact.
2. ***Bulk Mailing:*** The email which is sent in the form of bulk to many users.
3. ***Nameless Mails:*** The emails in which the identity of the user is not shown or is hidden.

3.3 Types of Spam:
There are various types of spam emails are present around the world, which are harmful to the emails. They are as follows

1. ***Email Phishing:*** Email Phishing is one of the most common ways of carrying out spam attacks on senders, and achieved through manipulating data. It is medium to scam the users entering into the personal information through fake Web sites using email forged and look like it is came from bank or any organization like PayPal.

2. ***Appending:*** The marketer have database in which contain names, addresses and telephone number of each customer, who will pay to have their database matched against a database containing email addresses and purchasing a list of e-mail addresses that match a list of those names to be used later become spam.

3. ***Image Spam:*** Which is Image spam is an obfuscating method in which the text of the message is stored as a GIF or JPEG image and displayed in the e-mail. This prevents text based spam filters from detecting and blocking spam messages. Image spam is currently used largely to advertise stocks etc.

4. ***419 Scams:*** Advance fee fraud spam such as the Nigerian "419" scam may be sent by a single individual from a cyber cafe in a developing country, in which individual receiving such spam could believe in it and would be scammed to give away money or do certain illegal things on behalf of them without him knowing so.

5. ***Blank Spam:*** Blank spam is spam lacking a payload advertisement that often the message body is missing along with subject line. Still, it fits the definition of spam because of its nature as bulk and unsolicited e-mail.

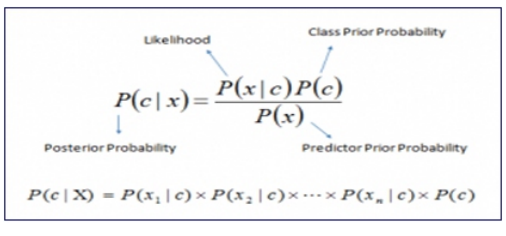
   Blank spam may be originated in different ways, either intentional or unintentionally. Where blank spam have sent in a directory harvest attack, to gather valid addresses from an e-mail service provider. The goal of this attack is to bounce to separate invalid addresses from the valid ones. In addition some spam may appear to be blank when in fact it is not present. An example of this is the VBS. Davina. B e-mail worm which propagates through messages that have no subject line and appears blank, when in fact it uses HTML code to download other files also.

6. ***Backscatter Spam:*** Backscatter is a side-effect of e-mail spam, where e-mail servers receiving spam and other mail send bounce messages to an innocent party. This occurs because the original message's envelope sender is forged to contain the e-mail address of the victim. A very large proportion of such e-mail is sent with a forged From: header, matching the envelope sender.

3.4 Email Classification Algorithms for Spam detection:
In this study I used three different email classification algorithms for spam detection .They are described as follows:

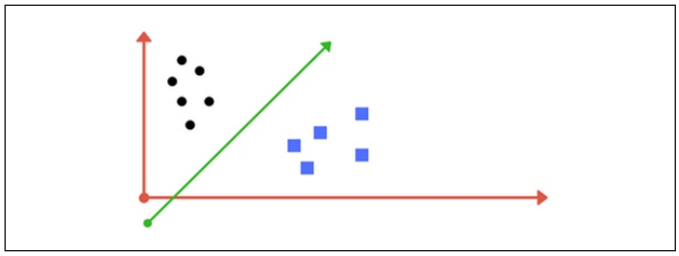
1. ***Naive Bayes Classification Algorithm:*** Which is a classification technique based on Bayes' Theorem that assumes independence of among predictors? Naive Bayes classifier states that "presence of one particular feature in a class is unrelated to the presence of any other feature classes". Which is made easy to simplify problems in the computations involved, hence it is called "naive". This classifier is also called idiot Bayes, simple Bayes, or independent Bayes.

   Bayes theorem provides a way to calculate posterior probability *P(c|x)* from *P(c)*, *P(x)* and *P(x|c)* and equation is as below:



   In which,
   - *P(c|x)* is the posterior probability of the class (*c*, target) which give predictor (*x*, attributes).
   - *P(c)*,which is the prior probability of the class.
   - *P(x|c)* ,which is likeliyhood probability of predictor given class.
   - *P(x)*, which is the prior probability of predictor.

2. ***Support Vector Machine Classifier Algorithm:*** Support Vector Machine (SVM) is a discriminative classifier which is defined by a separating hyper plane. In two dimensional space, hyper plane is a line which divides a plane in two parts where each class lay inside either outside of the line. The following diagram is shown as follows.



   If there is only two classes of dataset then it is called as a Binary SVM Classifier. Mainly there are two different types of SVM classifiers they are:

   1. Linear SVM Classifier

   2. Non-Linear SVM Classifier

   1. ***Linear Classifiers:*** Separating the data points in linear order by using a hyper-plane is classified as linear classifiers. There are different hyper-plane but the best way to separate the data using hyper-plane is by maximum margin difference viz. the distance of hyper-plane and the closest information point of any class.

   2. ***Non-Linear Classifiers:*** In Some cases the data is not separated properly or linearly in high dimensional plane for such separation non-linear classifiers are used which correctly classify the information points and label them to their exact class by using kernel tricks. Some mostly used kernel tricks are as Follows:

      a. *Homogenous kernels:* Polynomial kernels that are used for analysing the similarity of vectors are Represented by the expression below:

      $k(\vec{ai},\vec{aj})=(\vec{ai},\vec{aj})d$

      Where k is the kernel function and $(\vec{ai},\vec{aj})$ are the vectors of the work space with d as the degree of the polynomial.

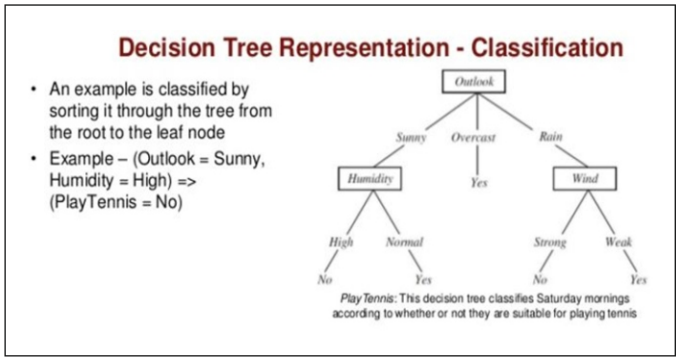
      b. *Non-Homogenous kernels:* In Non-homogenous kernels a free parameter is added that leverage the group of features combined together.

      $k(a,b)=(a^{T}b+c)d$

   3. ***C.J48 Classifier Algorithm:*** C4.5 algorithm known as j48 algorithm in WEKA. Which is used to generate a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan's earlier called as ID3 algorithm. The decision trees generated by C4.5 which is used for classification, so it is referred to as a statistical classifier and ranked #1 in the Top 10 Algorithms in Data Mining.

      A decision tree is a graph that uses a branching method to illustrate every possible outcome of the decision. Programmatically, this can be used to assign for monetary/time or other values to possible outcomes, so that decisions can be automated.

      Following is example decision tree.



4. EXPERIMENTS SETUP:
In this study, an analysis report on experimental evaluation of the three algorithms is presented as a result. Primarily the dataset which contains the spam

email used for evaluation performance is described in deep and short description of the various measures of formulas is outlined. The final result analysis of experimental results are discussed in details and presented as of bar chart and table formats.

a. *WEKA:* WEKA means (Waikato Environment for Knowledge Analysis), which is an innovation tool in data mining and machine learning research communities environment. This is developed by WEKA team since 1994 and contains many inbuilt algorithms for data mining and machine learning systems. It is open source and user friendly with freely available platform-independent software machine learning system. User who are not familiar and doesn't have knowledge about data mining can also use this software very easily, as it provides flexible facilities for scripting and experiments evaluations. As on new algorithms appear in research literature are updated in software and avail for users.

b. *Steps to Execute:* The steps to perform using data mining in WEKA is as follows:

- Data pre-processing and visualization
- Attribute selection
- Classification (Decision trees)
- Prediction (Nearest neighbour)
- Model evaluation
- Clustering (Cobweb, K-means)
- Association rules

c. *Datasets:* In this research the dataset for spam email is used, which is present for publicly from the Network Repository (http://networkrepository.com), i.e. SPAM E-mail Database. The dataset contains 57 attributes and 4601insteances in which 1813 emails are spam and the remaining are 2788 not spam emails. The dataset data type is multivariate with real and integer attributes of values.

d. *Performance Evaluation:* In this section the performance is evaluated by using measurement parameters for email classification comparison to data mining algorithms Naive Bayes, SVM and j48.

The measurement parameters is as follows:

1. **Accuracy:** The ratio of True Positive and True Negative combined against the total number of instances analysed.

   True Positives (TP) – which is the correctly predicted positive values which means that the value of actual is correct and the value of predicted is also correct.

   True Negatives (TN) – which is the correctly predicted negative values which means that the value of actual is not correct and value of predicted is also not correct.

   False positives and false negatives, these values occur when actual values are in contradicts with the predicted values.

   False Positives (FP) – When actual value is not correct and predicted value is correct.

   False Negatives (FN) – When actual value is correct but predicted value in not correct.

$$Accuracy = \frac{TP + TN}{TN + TP + FN + FP}$$

2. **Precision:** It determines the percentage of identification that are actually correct out of the total amount of cases claimed as correct.

$$Precision\,(P) = \frac{TP}{FP + TP}$$

3. **Recall:** It determines the percentage of identification that are actually correct out of the total amount of cases that should have been identified as correct.

$$Recall\,(R) = \frac{TP}{TP + FN}$$

5. RESULTS AND DISCUSSION:

By using the Spambase dataset as an input to the algorithms Naive Bayes, SVM and j48 and run in the WEKA environment and the final results are as follows.

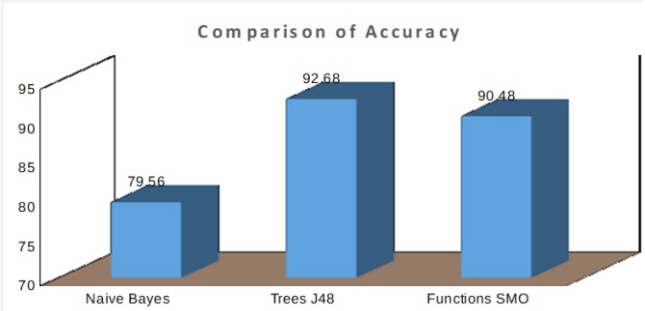
The total result is divided in 2 categories in the class column by indicating Spam as 1 and No Spam as 0.

Table 1: A comparison of results between three algorithms

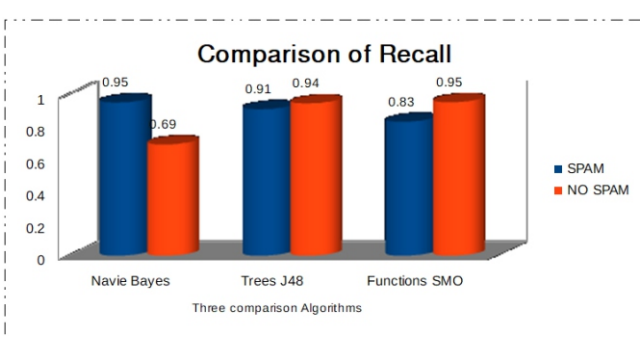
Classification algorithm	Accuracy (TP+TN)/ (TP+TN+FP+FN)	Recall TP/(TP+FN)	Precision TP/(TP+FP)	FP RATE FP/(FP+TN)	TP RATE (=RECALL)	F MEASURE 2*PR/(P+R)
NAÏVE BAYES	79.56	0.951	0.666	0.310	0.951	0.784
		0.690	0.956	0.049	0.690	0.801
J48	92.68	0.908	0.913	0.056	0.908	0.911
		0.944	0.940	0.092	0.944	0.942
SVM	90.48	0.831	0.918	0.048	0.831	0.873
		0.952	0.897	0.169	0.952	0.923

Classification Algorithm	Confusion Matrix		
	a	b	Classified as
NAIVE BAYES	1725	88	a=1 b=0
	865	1923	
J48	1646	167	a=1 b=0
	156	2632	
Functions SMO	1507	306	a=1 b=0
	134	2654	

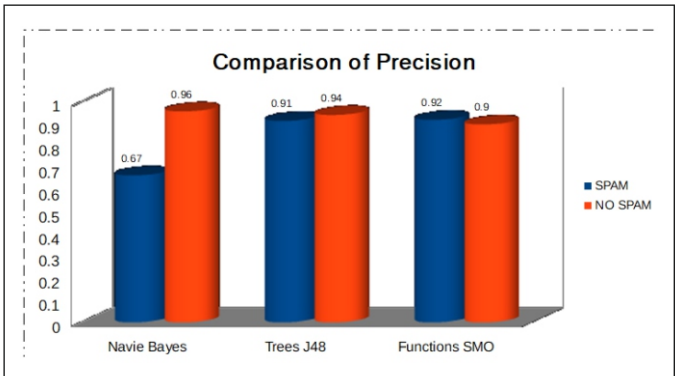
1. **Accuracy:** In this comparison of bar chat the Decision tree J48 algorithm is showing highest accuracy rate compare to Function SMO and Naive Bayes. It shows threat J48 algorithm is best for detection of Spam email.



2. **Recall:** In this comparison of bar chat the Naive Bayes algorithm is showing highest Recall rate for spam compare to J48 and Function SMO.



3. **Precision:** In this comparison of bar chat the Function SMO algorithm is showing highest Precision rate for Spam compare to J48 and Naive Bayes.



According to the comparative on three algorithms, namely Navie Bayes, J48, Function SMO to determine the best Email classification to prevent spam, and no spam in data mining. These algorithms utilized the spam email dataset, which is available publicly from the Network Repository in the WEKA environment. The common measures for the research is evaluated with respect parameters are Accuracy, Precision, and Recall.

Features	Ranking			Categories
	1	2	3	
Accuracy	J48	SVM	NAÏVE BAYES	
Precision	SVM	J48	NAÏVE BAYES	Spam
	NAÏVE BAYES	J48	SVM	No Spam
Recall	NAÏVE BAYES	J48	SVM	Spam
	SVM	J48	NAÏVE BAYES	No Spam

In above table it indicates that J48 is the best algorithm in terms of accuracy and also performs better in Recall and Precision

J48 creates decision trees from a labelled data and utilized to take a decision by dividing the data as reduced subsets for this study. The normalized data added information or entropy variation during splitting is done. The decision is taken based on the maximum normalization by gain of attributes to process.

The ability of J48 decision trees is used for missing values, value ranges, etc. Which makes it a superior algorithm compared to other algorithms. In this research it is observed that no algorithm shows 100% accuracy for finding spam in Email classification.

6. CONCLUSION AND FUTURE STUDY:

This paper present a method to classify mails based on three classifiers, i.e. J48, SVM, and Naïve Bayes. This classifiers were evaluated to separate spam from the email dataset by using WEKA tool kit. The emails were identified as spam (1) or not spam (0), which reflected the attributes of the dataset of email for spam filtering.

The algorithm was checked against parameters such as Accuracy, Precision, Recall. The analysis of the results demonstrated clearly that even though J48 is a very simple classifier which uses a decision tree, it gave the most accurate result in the experiment (92.68%).

SVM also present good results with accuracy of 90.48% and better performance results in other parameters too. But Naive Bayes is given accuracy of (79.56%), which is poor results in comparison to other classification.

In further research study it is required to improve an depth analysis algorithmlike Genetic algorithm, and classification techniques for finding the spam. In addition, different algorithms which are not included in WEKA should be added to test and experiments with various feature of attributes selections for comparisons.

REFERENCES:

1. Harjot Kaur "Survey on e-mail spam detection using supervised approach with feature selection", International Journal of Engineering Sciences & Research Technology (IJESRT), ISSN: 2277-9655, Impact Factor: 4.116, April, 2017
2. Deepika Mallampati, Nagaratna P. Hegde "A Machine Learning Based Email Spam Classification Framework Model: Related Challenges and Issues" International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278 -3075, Volume-9, Issue-4, February 2020.
3. Asif Karim "A Comprehensive Survey for Intelligent Spam Email Detection" IEEE, Volume: 7, 20 November 2019.
4. Anis Ismail, Shadi Khawandi, Firas Abdallah "Image Spam Detection: Problem and Existing Solution", International Research Journal of Engineering and Technology (IRJET), Volume: 06, e-ISSN: 2395-0056, Issue: 02 | Feb 2019.
5. Balogun Abiodun Kamoru "Spam Detection Approaches and Strategies: A Phenomenon", International Journal of Applied Information Systems (IJAIS) – ISSN: 2249-0868, Volume 12 – No. 9, December 2017.
6. Muhammad Iqbal, "Study on the Effectiveness of Spam Detection Technologies", I.J. Information Technology and Computer Science, Jan 2016, 01, 11-21.
7. Dipak R. Kawade, "SMS Spam Classification using WEKA", International Journal of Electronics Communication and Computer Technology (IJECCT) Volume 5 Issue ICICC (May 2015) ISSN: 2249-7838.
8. Madhvi Sharma, "A Survey of Email Spam Filtering Methods", ISSN 2224-5774 (Paper) ISSN 2225-0492 (Online) Vol.7, 2018 www.iiste.org
9. M. K. Paswan, P. S. Bala, G. Aghila, "Spam filtering: Comparative analysis of filtering techniques", Proc. Int. Conf. Adv. Eng. Sci. Manage. (ICAESM), pp. 170-176, Mar. 2012.
10. WEKA at "http://www.cs.waikato.ac.nz/~ml/weka"
11. http://networkrepository.com
12. Gaganjot Kaur,Amit Chhabra,"Improved J48 Classification Algorithm for the Prediction of Diabetes", International Journal of Computer Applications (0975 – 8887) Volume 98 – No.22, July 2014.